**Hubei University of Technology**

Big Data Analysis Homework Report

Class: computer science Name: 颜如玉 - 2221551212

Tutor: Prof. Jianxia Chen

Date: No: 4-（1）

Assignment： **Statistics and Probability**

1. **(20 points)Explain the following concepts and give an example for each concept.**
2. **Central limit theorem**

The mean of a large number of independently and identically distributed (iid) random variables (with mean μ and standard deviation σ) is approximately normally distributed, with mean μ and standard deviation σ/sqrt(n), where n is the sample size.

1. **standard error of the mean**

The standard error of the mean is a method used to determine the standard deviation of a sampling distribution provided for a population. It measures the extent by which the sample mean or average differs from the true population mean. The standard error of the mean gives an accurate idea of the sample mean by analyzing the various samples from the sampling distribution.

1. **confidence interval**

The confidence intervalis the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re-sample the population in the same way.

Standard Error of the Mean (SEM): standard deviation / sqrt(n)

95% confidence interval: 1.96\*SEM (the interval that is likely to include the real mean with 95% probability)

In [965]:

measures = randint(0, 100, size=(10,3))

In [966]: measures

Out[966]: array([[70, 54, 67], [62, 24, 60], [ 0, 61, 11], ..., [78, 43, 94], [45, 79, 81], [54, 50, 29]])

1. **z-score**

A z-score describes the position of a raw score in terms of its distance from the mean when measured in standard deviation units. The z-score is positive if the value lies above the mean and negative if it lies below the mean.

z-score formula

• Zi = (xi - μ) / σ

• Z-score is unit-less, can be + or –

• When distribution is approx. normal, z-score can be conveniently mapped to probabilities

1. **Correlation**
2. **PMF**
3. **CDF**
4. **PDF**
5. **Probability (40 points). Show the results and your calculation using python-style code**
6. Toss a fair coin 5 times, what is the probability of seeing 5 heads in a row?

P(HHHHH | fair coin) = (0.5)\*\*5 = 0.0313

1. Given a box that contains 90% fair coins and 10% loaded coins, (a loaded coin gives heads 90% of the time), what is the probability for a randomly drawn coin to give 5 heads in a row?

P(HHHHH | random coin from box of mixed coins) =

1. Given a coin randomly drawn from the box mentioned above in 2b, what is the probability to get exactly 9 heads in 10 tosses?

P(9 H and 1T | random coin from box of mixed coins) =

1. If you randomly pick a coin from the box mentioned above in 2b, toss it 5 times and get all heads. What is the probability that this is a loaded coin?

P(loaded | HHHHH) =

1. If you randomly pick a coin from the box mentioned above in 2b, toss it 10 times and get nine heads and one tail. What is the probability that this is a loaded coin?

P(loaded | 9H and 1T) =

1. **Statistics (40 points)**
2. Write a program to simulate tossing a fair coin for 100 times and count the number of heads. Repeat this simulation 10\*\*5 times to obtain a distribution of the head count and plot the histogram as well as CDF. Label your plots clearly.
3. Use the binomial distribution CDF (use scipy.stats.binom.cdf) to estimate the probability of having NO MORE THAN k heads out of 100 tosses, where k = 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Do these probabilities agree with the numbers of head counts you obtained in 1a? (Plot the head counts you obtained from the simulation results in 3a against the probabilities from your theoretical calculation here. Plot in loglog scale is probably needed to visualize small probabilities.)
4. Make a normal probability plot (thinkstats ch 4.4) to show that this distribution is close to a normal distribution with mean 50 and standard deviation 5.
5. Use normal distribution approximation to calculate the cumulative probabilities that you were asked to calculate in 3b, and compare the two results using a loglog plot. (Hint: If head count follows a normal distribution with mean = 50 and std = 5, a head count of 40 is equivalent to z-score = -2, and the corresponding CDF can be calculated using scipy.stats.norm.cdf.)